Atty. Dkt. No.: 035905/0118

U.S. PATENT APPLICATION

For

TWO MASK FLOATING GATE EEPROM AND METHOD OF MAKING

Inventors: Igor G. Kouznetsov and Andrew J. Walker

Prepared by:
Foley & Lardner
3000 K St. N.W.
Washington D.C. 20007
(202) 672-5300

TWO MASK FLOATING GATE EEPROM AND METHOD OF MAKING

[0001] This application is a continuation-in-part of U.S. Application Serial Number 09/927,648, filed on August 13, 2001, which is incorporated by reference in its entirety. This application also claims benefit of priority of provisional application 60/279,855 filed on March 28, 2001, which is incorporated by reference in its entirety.

FIELD OF THE INVENTION

[0002] The present invention is directed generally to semiconductor devices and methods of fabrication and more particularly to a nonvolatile EEPROM memory device and method of fabrication.

BACKGROUND OF THE INVENTION

[0003] U.S. Patent No. 5,768,192, issued to B. Eitan, and the technical article entitled "NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell" by B. Eitan et al. in *IEEE Electron Device Letters*, vol. 21, No. 11, Nov. 2000, pp. 543-545 teach a nonvolatile semiconductor memory cell which uses asymmetrical charge trapping in the nitride charge storage layer of the Oxide-Nitride-Oxide (ONO) stack to store two bits in one cell. The cell is written by hot electron injection into the charge storage layer above the drain junction. The cell is read in the opposite direction to which it was written, i.e., voltages are applied to the source and gate, with the drain grounded. The memory cell is constructed in a p-type silicon substrate. However, this silicon-oxide-nitride-oxide-silicon (SONOS) 1TC memory requires LOCOS (localized oxidation of silicon) isolation regions, which cause the cell area to be larger than desirable, and leads to a less than optimum cell density and increases the number of photolithographic masking steps.

[0004] Another type of prior art memory device is disclosed in the technical article entitled "A Novel Cell Structure for Giga-bit EPROMs and Flash Memories Using Polysilicon Thin Film Transistors" by S. Koyama in *1992 Symposium on VLSI Technology Digest of Technical Papers, pp. 44-45*. As shown in Figure 1, each memory cell is a "self-aligned" floating gate cell and contains a polycrystalline silicon thin film transistor electrically erasable programmable read only memory (TFT EEPROM) over an insulating layer. In this device, the bit lines extend in the direction parallel to the source-channel-drain direction (i.e., the bit lines extend parallel to the charge carrier flow direction). The word lines extend in the direction perpendicular to the source-channel-drain direction (i.e., the word lines extend perpendicular to the charge carrier flow direction). The TFT EEPROMs do not contain a separate control gate. Instead, the word line acts as a control gate in regions where it overlies the floating gates.

[0005] The layout of Koyama requires two polycide contact pads to be formed to contact the source and drain regions of each TFT. The bit lines are formed above the word lines and contact the contact pads through contact vias in an interlayer insulating layer which separates the bits lines from the word lines. Therefore, each cell in this layout is not fully aligned, because the contact pads and the contact vias are each patterned using a non-self-aligned photolithography step. Therefore, each memory cell has an area that is larger than desirable, and leads to a less than optimum cell density. The memory cell of Koyama is also complex to fabricate because it requires the formation of contact pads and bit line contact vias, which requires separate photolithographic masking steps. Furthermore, the manufacturability of the device of Koyama is less than optimum because both bit lines and word lines have a non-planar top surface due to the non-planar underlying topography. This may lead to open circuits in the bit and word lines.

BRIEF SUMMARY OF THE INVENTION

[0006] A preferred embodiment of the present invention provides a floating gate transistor, comprising a channel island region, a source region located adjacent to a first side of the channel island region, a drain region located adjacent to a second side of the channel island region, a tunneling dielectric located above the channel island region and a floating gate having a first, second, third and fourth side surfaces, wherein the floating gate is located above the tunneling dielectric. The transistor also comprises a control gate dielectric located above the floating gate and a control gate located above the control gate dielectric. The first and second side surfaces of the control gate are aligned to third and fourth side surfaces of the channel island region, and to third and the fourth side surfaces of the floating gate.

[0007] Another preferred embodiment of the present invention provides a method of making a floating gate transistor, comprising providing a semiconductor active area, forming a tunnel dielectric layer over the active area, forming a floating gate layer over the tunnel dielectric layer, forming a first photoresist mask over the floating gate layer, patterning the floating gate layer using the first photoresist mask to form a floating gate rail and doping the active area using the floating gate rail as a mask to form source and drain regions in the active area. The method further comprises forming an intergate insulating layer adjacent to lower portions of side surfaces of the floating gate rail, forming a control gate dielectric layer over and adjacent to upper portions of the side surfaces of the floating gate rail, forming a control gate layer over the control gate dielectric layer, forming a second photoresist mask over the control gate layer, and patterning the control gate layer, the control gate dielectric layer, the floating gate rail, the tunnel dielectric layer and the active area using the second photoresist mask to form a control gate, a

control gate dielectric, a floating gate, a tunnel dielectric and a channel island region.

[0008] Another preferred embodiment of the present invention provides a method of forming an array of floating gate transistors, comprising forming at least portions of a plurality of floating gates over a tunnel dielectric located over a semiconductor active area, doping the active area using the at least portions of the plurality of floating gates as a mask to form a plurality of bit lines in the active area and forming an intergate insulating layer between lower portions of side surfaces of the at least portions of the plurality of floating gates. The method further comprises forming a control gate dielectric on exposed upper surfaces of the at least portions of the floating gates and on exposed upper portions of side surfaces of the at least portions of the floating gates, and forming a plurality of word lines over the control dielectric and over the intergate insulating layer.

[0009] Another preferred embodiment of the present invention provides method of making a floating gate transistor, comprising forming the entire floating gate transistor using two photolithographic masking steps.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] Figure 1 is diagram of the process flow of a method of making a prior art memory device.

[0011] Figure 2 is a side cross-sectional view of an in process memory array after the formation of the floating gate rails according to the first preferred embodiment of the present invention.

[0012] Figure 3 is a top view of Figure 2.

[0013] Figure 4 is a side cross-sectional view of an in process memory array after floating gate rail patterning and bit line implantation and silicidation according to the first preferred embodiment of the present invention. The cross-section is perpendicular to the bit lines.

[0014] Figure 5 is a top view of Figure 4.

[0015] Figure 6 is a side cross-sectional view of an in process memory array after the formation of the intergate insulating layer according to the first preferred embodiment of the present invention. The cross-section is perpendicular to the bit lines.

[0016] Figure 7 is a side cross-sectional view of the array after the formation of the control gate layer according to the first preferred embodiment of the present invention. The cross-section is perpendicular to the bit lines.

[0017] Figure 8 is a side cross-sectional view of the array after the patterning of the control gate layer according to the first preferred embodiment of the present invention. The cross-section is taken along line A-A' in Figure 7, and is parallel to the bit lines.

[0018] Figure 9 is a top view of Figure 8.

[0019] Figure 10 is a three dimensional view of Figures 8 and 9.

[0020] Figure 11 is a side cross-sectional view of the array according to the second preferred embodiment of the present invention. The cross-section is perpendicular to the bit lines.

[0021] Figures 12-13 are schematic side cross-sectional views of angled ion implantation methods to form asymmetric source and drain regions of the second preferred embodiment of the present invention.

[0022] Figure 14 is a schematic side cross-sectional view of a three dimensional array according to the third preferred embodiment of the present invention.

[0023] Figures 15-18 are schematic side cross-sectional views of formation of interconnects between device levels according to a fourth preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0024] The present inventors have realized that memory cell area is enlarged by misalignment tolerances that are put into place to guarantee complete overlap between features on different layers. Thus, the present inventors have developed a fully aligned memory cell structure which does not require misalignment tolerances. Therefore, such a cell structure has a smaller area per bit (i.e., per cell) and uses fewer mask steps. The fully aligned cell structure increases memory density and decreases die size and cost. Furthermore, by optionally stacking the cells vertically in the Z-direction, the memory density is further increased, which leads to further decreases in the die size and cost.

[0025] As described with respect to the preferred embodiments of the present invention, an entire floating gate EEPROM transistor may be made using only two photolithographic masking steps. This decreases the process cost and complexity and ensures the precise alignment or self-alignment of the layers of the transistor, since many of these layers are patterned together using the same photoresist mask.

[0026] For example, by patterning the control and floating gates using the same mask as the channel results in a channel island which has at least two side surfaces which are aligned to the floating and control

gates. Furthermore, forming channel regions as islands and then filling the trenches between the islands with an insulating layer creates trench isolation between adjacent transistors without requiring an extra photolithographic masking step. In contrast, an extra photolithographic masking step is required to form prior art LOCOS or trench isolation between transistors.

[0027] Furthermore, the separate photolithographic masking step to form bit lines of the prior art process of Figure 1 may be eliminated by forming the bit lines using the same masking step used to form the floating gates. For example, the floating gate layer is patterned only in one direction to form floating gate rails or strips. Then, the semiconductor active region portions exposed between the floating gate rails is implanted and/or silicided to form conductive rails or strips extending parallel to the floating gate rails. These conductive rails in the active region act as the bit lines. These bit lines are then covered by an intergate insulating layer which is formed in self-alignment between the floating gate rails. The floating gate rails are then patterned to form discrete floating gates using the same mask as is used to form the channel regions and the control gate. Since the intergate insulating layer covers the bit lines during this etching step, the bit lines are not etched during the channel etching step. The portions of the bit lines adjacent to the patterned floating gates act as the source and drain of a given EEPROM transistor.

[0028] In this configuration, bit line contact pads (i.e., source and drain electrodes) and bit line contact vias are not required because the bit lines may be formed in self-alignment with the EEPROM floating gates. Furthermore, since the EEPROMs are fully aligned or self-aligned, the bit and word lines may have a substantially planar upper surface, which improves the reliability of the device.

[0029] The method of making the array of EEPROM transistors 1 according to the first preferred embodiment of the present invention will now be described in detail with references to Figures 2-10. It should be noted that the present invention is not limited to an array of transistors, and includes the formation of a single transistor. It should also be noted that the array 1 does not have to be formed in a silicon layer located on an insulating surface (i.e., does not have to be formed as a TFT array), and may instead be formed in a bulk silicon substrate to form a bulk MOSFET EEPROM array.

[0030] A substrate having an insulating surface (i.e., a Silicon-On-Insulator (SOI) substrate) is provided for the formation of the memory array. The substrate may comprise a semiconductor (i.e., silicon, GaAs, etc.) wafer covered with an insulating layer, such as a silicon oxide or nitride layer, a glass substrate, a plastic substrate, or a ceramic substrate. In a preferred aspect of the first embodiment, the substrate is a monocrystalline bulk silicon substrate that has received prior processing steps, such as forming CMOS (complementary metal oxide semiconductor) transistors in the substrate. The CMOS transistors may comprise peripheral or driver circuitry for the memory array. In the most preferred aspect, the circuitry comprises row and column address decoders, column input/outputs (I/O's), and other logic circuitry. However, if desired, the driver circuitry may be formed on an insulating substrate, such as a silicon-on-insulator substrate, a glass substrate, a plastic substrate, or a ceramic substrate. The silicon-on-insulator substrate may be formed by any conventional method, such as wafer bonding, Separation by Implantation of Oxygen (SIMOX), and formation of an insulating layer on a silicon substrate. After the peripheral circuitry is completed, an interlayer insulating layer (also known as an interlayer dielectric) 3 is conformally deposited over the circuitry as shown in Figure

2. The interlayer insulating layer 3 may comprise one or more of any suitable insulating layers, such as silicon oxide, silicon nitride, silicon oxynitride, PSG, BPSG, BSG, spin-on glass and/or a polymer dielectric layer (such as polyimide, etc.). The interlayer insulating layer 3 is preferably planarized using chemical-mechanical polishing (CMP), etch back and/or any other means.

[0031] A semiconductor active area layer 5 is then deposited over the insulating layer 3 to complete the SOI substrate. The semiconductor layer will be used for the transistor active areas. Layer 5 may have any desired thickness, such as 10 to 120 nm, preferably less than 100 nm, most preferably less than 30 nm. Layer 5 is chosen so that in depletion regime the space charge region below the transistor gate extends over the entire layer. The layer 5 may be thinned to the desired thickness using wet silicon etching or by sacrificial oxidation following by a wet oxide etch. Preferably, the semiconductor layer 5 comprises an amorphous or polycrystalline silicon layer doped with first conductivity type dopants. For example, layer 5 may be p-type doped by in-situ doping during deposition, or after deposition by ion implantation or diffusion.

[0032] If desired, the crystallinity of the semiconductor layer 5 may be improved by heating the layer 5. In other words, an amorphous silicon layer may be recrystallized to form polysilicon or a grain size of a polysilicon layer may be increased. The heating may comprise thermal or laser annealing the layer 5. If desired, catalyst induced crystallization may be used to improve the crystallinity of layer 5. In this process, a catalyst element such as Ni, Ge, Mo, Co, Pt, Pd, a silicide thereof, or other transition metal elements, is placed in contact with the semiconductor layer 5. Then, the layer 5 is thermally and/or laser annealed. During the annealing, the catalyst element either propagates through the silicon layer leaving a trail of large grains, or serves as a seed where silicon

crystallization begins. In the latter case, the amorphous silicon layer then crystallizes laterally from this seed by means of solid phase crystallization (SPC).

[0033] It should be noted that the deposition of amorphous or polysilicon layer 5 may be omitted if a single crystal SOI substrate is used. In this case, using the SIMOX method, oxygen ions are implanted deep into a single crystal silicon substrate, forming a buried silicon oxide layer therein. A single crystal silicon layer remains above the buried silicon oxide layer.

[0034] Next, the surface of the active area layer 5 is preferably cleaned from impurities and a native oxide is removed. A tunnel dielectric 7 is formed on the active area layer 5. Preferably, the tunnel dielectric comprises a thermally grown silicon oxide layer (i.e., a silicon dioxide layer grown on the silicon layer 5 by exposing layer 5 to an oxygen containing atmosphere to convert a top portion of layer 5 to silicon dioxide). The tunnel dielectric has a thickness of 5 nm to 10 nm, preferably 7 nm. It should be noted that different layer thicknesses and different materials, such as silicon nitride or silicon oxynitride, may be used instead.

[0035] After the tunnel dielectric 7 is formed, a floating gate layer 9 is deposited over, and preferably directly on the tunnel dielectric 7. The floating gate layer 9 preferably comprises a polysilicon layer, such as an N+ polysilicon layer. Such a polysilicon layer may have any appropriate thickness, such as 100 to 300 nm, preferably 200 nm, and any appropriate dopant concentration, such as 10^{19} - 10^{21} cm^{-3}, preferably 10^{20} cm^{-3}.

[0036] If desired, an optional hardmask or etch stop layer 11, such as a silicon oxide layer or a dual lower silicon oxide / upper silicon nitride

film, is formed on the surface of the floating gate layer 9. Layer 11 may have any appropriate thickness, such as, for example 20-200 nm, preferably 50 nm. Materials other than silicon oxide and silicon nitride may be used for layer 11, if desired.

[0037] Next, a bit line pattern is transferred to the in process array using a reverse bit line mask, as shown in Figure 2. For example, a positive photoresist layer 13 is formed over the hardmask layer 11 and then exposed through the reverse bit line mask and developed. Of course, if a negative photoresist is used, then the clear and the opaque areas of the bit line mask are reversed.

[0038] The photoresist mask 13 features are etched into the hardmask layer 11, and the floating gate layer 9, to form a plurality of rail stacks 15, as shown in Figure 2. The tunnel dielectric 7 serves as an etch stop layer. Then, the photoresist mask 13 is stripped from the patterned gate rail stacks 15.

[0039] Figure 3 illustrates the top view of the in process array shown in Figure 2. As shown in Figure 3, the rail stacks 15 are in the shape of strips, and contain the floating gate rails 9 and the hardmask rails 11. The tunnel dielectric 7 is exposed in the areas between the rail stacks 15. If desired, an optional thin layer of silicon nitride, oxynitiride or oxide is grown to seal the exposed sidewalls of the floating gate rails 9.

[0040] The array bit lines 17 are formed by self-aligned ion implantation into the active layer 5, using the rail stacks 15 as a mask, as shown in Figure 4. The photoresist layer 13 is removed prior to the implantation. Alternatively, it may be left on the rail stacks 15 during this implantation. The ion implantation is carried out through the tunnel dielectric 7. However, if desired, the portions of the tunnel dielectric 7

between the floating gate rails 9 may be removed prior to the ion implantation.

[0041] Channel regions 19 in the active layer 5 are located below the floating gate rails 9. The bit lines 17 are doped with a second conductivity type dopant different from the first conductivity type dopant of the channels 19. Thus, if the channels 19 are p-type doped, then the bit lines 17 are n-type doped, and vice-versa.

[0042] Next, optional sidewall spacers 21 are formed on the sidewalls of the rail stacks 15, as shown in Figure 4. Preferably, the spacers 21 comprise silicon oxide or silicon nitride. Most preferably, the spacers comprise a different material from the hardmask layer. The spacers 21 are preferably formed by conformal deposition of a silicon oxide layer over the stacks 15, followed by an anisotropic oxide etch. The spacer etch process concludes with an etch process for the tunnel dielectric 7 to expose the bit lines 17. Doping in the bit lines 17 may be increased at this time by additional self-aligned ion implantation, using the rail stacks 15 and spacers 21 as a mask, if desired. In this case, the implantation before spacer formation is used to form lightly doped portions or extensions of the source/drain (LDD) portions (about $1x10^{16}$ to about $1x10^{18}$ cm^{-3} doping concentration) while the doping after spacer formation is used to form heavily doped source and drain regions (about $1x10^{19}$ to about $1x10^{21}$ cm^{-3} doping concentration). Preferably, the bit lines a n-type doped. However, p-type doping may be used instead. The formation of the spacers 21 and the lightly doped extensions may be omitted if desired.

[0043] The salicide process is then used to form silicide regions 23 in the top of the bit lines 17 in a self-aligned fashion, as shown in Figure 4. The salicide process comprises three steps. First a layer of metal, such

as Ti, W, Mo, Ta, etc., or a transition metal such as Co, Ni, Pt or Pd is blanket deposited over the exposed bit line regions 17, the sidewall spacers 21 and the hardmask layer 11 of the rail stacks 15. The array is annealed to perform a silicidation by direct metallurgical reaction, where the metal layer reacts with the silicon in regions 17 to form the silicide regions 23 over regions 17. The unnreacted metal remaining on the spacers 21 and the hardmask layer 11 is removed by a selective etch, e.g., by a piranha solution. The silicide regions 23 comprise portions of the bit lines 17 in addition to the previously doped silicon regions in the active layer 5.

[0044] Figure 5 shows the top view of the device in Figure 4 at this stage in the processing. The bit lines 17 containing silicide regions 23 extend as strips parallel to the rail stacks 15.

[0045] A conformal intergate insulating layer 25 is then deposited to fill the trenches above the bit lines 17 and between the floating gate rails 15 and sidewall spacers, as shown in Figure 6 (the sidewall spacers 21 are merged into layer 25 in Figure 6). The insulating layer 25 may comprise any insulating material, such as silicon oxide, silicon oxynitride, phosphosilicate glass (PSG), borophosphosilicate glass (BPSG), borosilicate glass (BSG), spin-on glass, a polymer dielectric layer (such as polyimide, etc.), and/or any other desired insulating material. Preferably, the intergate insulating layer is an isolation silicon oxide layer deposited by a high density plasma (HDP) method.

[0046] The intergate insulating layer 25 is formed over and between the rail stacks 15 (i.e., over and adjacent to the floating gate rails 9). The intergate insulating layer 25 is then etched back such that the intergate insulating layer 25 remains adjacent to lower portions 27 of the floating gate rail side surfaces, below top portions 29 of the floating gate rail side

surfaces. Thus, the top portions 29 of the side surfaces of the floating gate rails 9 are exposed during this etchback. Furthermore, the top portions of the sidewall spacers 21 and the hardmask layer 11 are removed from the floating gate rails 9 during the etchback, as shown in Figure 6.

[0047] A control gate dielectric layer 31 (also known as an inter-poly dielectric) is formed over the floating gate rails 9, as shown in Figure 7. The control gate dielectric layer 31 is formed on the upper portions 29 of the side surfaces of the floating gate rail 9 and on the top surface of the floating gate rail 9. The control gate dielectric layer 31 may have any appropriate thickness, such as 8 to 20 nm, preferably 12 nm of silicon oxide equivalent. The control gate dielectric may comprise several dielectric materials with varying dielectric permittivities. The silicon oxide equivalent is silicon oxide of such thickness that when used as a dielectric in a capacitor, it yields the same capacitance per unit area as the control gate dielectric. The control gate dielectric layer 31 may be grown on the control gate by thermal oxidation or deposited by CVD or other means. The control gate dielectric may comprise silicon oxide, silicon nitride, silicon oxynitride, or a stack comprising a thermally grown silicon oxide layer, a LPCVD deposited silicon nitride layer and a high temperature LPCVD deposited silicon oxide (HTO) layer.

[0048] If desired, the top surface and the upper portions 29 of the side surfaces of the floating gate rails 9 may be roughened prior to forming the control gate dielectric layer. The roughening may be accomplished by etching the exposed polysilicon of the rails 9 with an etching medium which selectively attacks polysilicon grain boundaries, such as wet etching with NH_3OH.

[0049] The control gate layer 33 is then deposited over the entire device, as shown in Figure 7. The control gate layer 33 is formed on the control gate dielectric layer 31 such that the control gate layer 33 is located over the top surface of the floating gate rails 9 and laterally adjacent to the upper portions 29 of the side surfaces of the floating gate rails. Since the control gate is located adjacent to the top and sides of the floating gate, this increases the capacitance between the floating and control gates. Preferably, the control gate layer 33 comprises a multilayer stack comprising a first N+ polysilicon layer, a silicide layer (such as a TiSi or WSi, etc.) and a second N+ polysilicon layer. The polysilicon layers are preferably 100-300 nm thick, such as 200 nm thick. The silicide layer is preferably 50 to 100 nm thick, such as 60 nm thick. The lower polysilicon layer fills in the openings between the floating gates and overlies the control gate dielectric. Alternatively, the control gate layer can also be a single layer of silicide, metal, or any other combination of heavily doped amorphous or polycrystalline silicon, silicide, and/or metal.

[0050] Next, a second photoresist mask 35 is formed by applying a photoresist layer over the control gate layer 33, exposing it through the word line mask and developing it. The second photoresist mask 35 is used as a mask to anisotropically etch the control gate layer 33, the control gate dielectric layer 31, the floating gate rails 9, the tunnel dielectric layer 7 and the active layer 5 to form a plurality of control gates 43, a plurality of control gate dielectrics 41, a plurality of floating gates 49, a plurality of tunnel dielectrics 47 and a plurality of channel island regions 19, as shown in Figure 8. Photoresist mask 35 is shown in dashed lines to indicate that it has already been removed in the array of Figure 8. Figure 8 is a cross sectional view along line A-A' in Figure 7.

[0051] Each control gate 43, control gate dielectric 41, floating gate 49, tunnel dielectric 47 and channel island region 19 comprise a second

rail stack 45 as shown in Figure 8. Thus, the sidewalls of the layers of the rail stacks 45 are aligned since the rail stacks 45 were formed during one etching step using the same mask 35.

[0052] If desired, the exposed sidewalls of the channel region islands 19, the floating gates 49 and the control gates 43 sidewalls may be optionally sealed by growing a thin layer of silicon nitride or oxide on them, for example by thermal nitridation or oxidation. This completes construction of the memory array 1. An insulating fill layer 50 is then deposited between the second rail stacks 45, and if necessary planarized by chemical mechanical polishing or etchback, over the control gates 43. Layer 50 acts as trench isolation fill between the adjacent channel island regions 19. Layer 50 preferably comprises the same material as the interlayer insulating layer 3.

[0053] Figure 9 is a top view and Figure 10 is a three dimensional view of memory array 1 after the second photoresist mask 35 is removed and layer 50 is formed. The array of EEPROMs shown in Figures 8-10 contains a plurality of bit line columns 17. Each bit line 17 contacts the source or the drain regions 57 of the TFT EEPROMs 51 (one exemplary memory cell or TFT EEPROM 51 is delineated by a dotted-dashed line in Figure 9). The source and drain regions 57 are portions of the bit lines 17 that are located adjacent to the floating gates 49. Thus, the bit lines and the source and drain regions are formed in the same step without requiring an extra photolithographic masking step. The columns of bit lines 17 extend substantially perpendicular to the source-channel-drain direction of the TFT EEPROMs 51 (i.e., at least a portion of the bit lines extends 0-20 degrees from this perpendicular direction). The bit lines 17 comprise rails which are located under the intergate insulating layer 25.

[0054] It should be noted that in a memory array 1, the designations "source" and "drain" are arbitrary. Thus, the regions 57 may be considered to be "sources" or "drains" depending on which bit line 17 a voltage is provided. Furthermore, since no field oxide regions are preferably used in this memory array, each region 57 is located between two floating gates 49. Therefore, a particular region 57 may be considered to be a "source" with respect to one adjacent floating gate 49, and a "drain" with respect to the other adjacent floating gate 49. A source region 57 is located adjacent to a first side of the channel island region 19, while a drain region 57 located adjacent to a second side of the channel island region 19, such that the channel region is located between the source and drain regions 57. Furthermore, the term "rail" and "rail stack" are not limited to strips which extend in only one direction, and the "rails" and "rail stacks" may have curves or bends and extend in more than one direction.

[0055] The array 1 also contains a plurality of word lines 53 which contain the control gates 43. In other words, the control gate 43 of each transistor comprises a portion of a word line 53. The rows of word lines 53 extend substantially parallel to the source-channel-drain direction of the TFT EEPROMs 51 (i.e., at least a portion of the word lines extends 0-20 degrees from this parallel direction).

[0056] The floating gates 49 comprise posts located between the channel islands 19 and the control gates 43. The posts 49 have four side surfaces as shown in Figures 6, 7, 8 and 9. The first 55 and second 56 side surfaces of the control gate 43 are aligned to third 59 and fourth 61 side surfaces of the channel island region 19, and to third 63 and the fourth 65 side surfaces of the floating gate 49, as shown in Figure 8. Furthermore, the first 55 and the second 56 side surfaces of the control

gate 43 are aligned to side surfaces of the control gate dielectric 41 and to side surfaces of the tunneling dielectric 47, as shown in Figure 8.

[0057] The word line photolithography step does not require misalignment tolerances, since the word lines 53 are patterned using the same mask as the floating gate rails 9 and the active layer 5 (i.e., channel regions 19) of each TFT 51 in the cell. Therefore, the word lines 53 are not only aligned to the floating gates 49 of the TFT EEPROMs 51 but are also aligned to the channel regions 19 of each memory cell. Furthermore, during the same etching step, the adjacent control gates, floating gates and channel islands are isolated from each other. By using a fully aligned memory cell, the number of expensive and time consuming photolithography steps are reduced. Furthermore, since no misalignment tolerances for each cell are required, the cell density is increased. Another advantage of the device of the first embodiment is that since a thick intergate insulating layer 25 is located between the bit lines 17 and the word lines 53, the parasitic capacitance and a chance of a short circuit between the bit lines and the word lines are decreased.

[0058] The nonvolatile memory array 1 may be programmed and erased by various conventional mechanisms. For example, the cells or TFT EEPROMs 51 of the array may be programmed or written by applying a programming voltage between the source and drain regions to achieve channel hot carrier (e.g., electron) injection into the floating gate 49. The cells or TFTs 51 of the array may be erased in blocks by applying an erase voltage between the control gate and a source or a drain to achieve Fowler-Nordheim carrier (i.e., electron) tunneling from the floating gate to the channel.

[0059] In a second preferred embodiment of the present invention, the array 100 contains cells or TFTs 151 which have asymmetric source

and drain regions 157, as shown in Figure 11. The drain overlap with the floating gate 149 is much larger than the source overlap with the floating gate 149. In Figure 11, a region 157 that acts as a source for one floating gate 149 acts as a drain for an adjacent floating gate 149. The other features of the TFTs 151 are the same as those of TFTs 51 described in the first preferred embodiment.

[0060] Due to the floating gate 149 to drain 157 overlap, the array 100 may be programmed or written bitwise by Fowler-Nordheim tunneling from the floating gate 149 to the drain 157. The control gate 143 is grounded, while a programming voltage is applied to the asymmetric drain regions 157, while the source regions 157 float. Since the source region is offset from the floating gate, no tunneling occurs from the floating gate to the source. This programming step decreases the threshold voltage of the TFT 151.

[0061] The cells or TFTs 151 of the array 100 may be erased in blocks by applying a high erase voltage to the control gate and grounding the source and drain regions to achieve Fowler-Nordheim carrier (i.e., electron) tunneling from the channel to the floating gate. This programming increases the threshold voltage of the TFT 151.

[0062] The asymmetric source and drain regions 157 may be formed by any desired method. For example, in one preferred method shown in Figure 12, the source and drain regions 157 are formed by performing an angled ion implant 161 using the first rail stack 115 as a mask. In Figure 12, the angled implant 161 comprises implanting heavily doped source and drain regions 157. If desired to form lightly doped portions of the source regions 163, then a lightly doped source region 163 is implanted at a smaller angle 165 sufficient to achieve an offset source as shown in Figure 13.

[0063] The first and second preferred embodiments describe and illustrate a cross-point array of word lines and bit lines at a horizontal level and a method of making thereof. Each memory cell consists of a single programmable field effect transistor (i.e., TFT), with its source and drain connected to the j^{th} bit line and the $(j+1)^{st}$ bit line, respectively, and a control gate being either connected to or comprising the k^{th} word line. This memory arrangement is known as the NOR Virtual Ground (NVG) Array (also referred to as VGA). If desired, the memory array may also be arranged in non volatile flash memory architectures other than VGA, such as NOR-type memory or Dual String NOR (DuSNOR) memory, for example. The DuSNOR architecture, where two adjacent cell strings share a common source line but use different drain lines, is described in K. S. Kim, et al., IEDM-95, (1995) page 263, incorporated herein by reference. The DuSNOR memory may be fabricated using the same process as the VGA memory, except that an additional masking step is used to pattern the active area layer 5 to separate the drain regions of adjacent cells. The active area 5 is patterned using a third mask to form a plurality of islands containing two EEPROM transistors sharing a common source.

[0064] Alternatively, this additional masking step can instead be used to separate both drains and sources of adjacent cells, thus achieving full isolation of adjacent cell strings. The active area 5 is patterned using a third mask to form a plurality of islands containing one EEPROM transistor to form an array with separated drain and source lines. Thus, each cell does not share a source or a drain (i.e., bit) line with a laterally adjacent cell. This NOR-type memory array architecture is known as Separated Source Line NOR (SSL-NOR) memory. The SSL-NOR array architecture is described in I. Fujiwara, et al., Proceedings of Non-Volatile Semiconductor Memory Workshop (2000), page 117, incorporated herein by reference.

[0065] The process sequence of the first and second preferred embodiments of the present invention requires only two photolithographic masking steps to form each cell. One masking step is for gate patterning / self aligned bit line formation. The other masking step is for word line patterning. The methods of the preferred embodiments of the present invention exploit self-alignment to reduce alignment tolerances between the masks. The memory cell area achieved with the foregoing process is about $4f^2$, where f is the minimum feature size (i.e. 0.18 microns in a 0.18 micron semiconductor process). The term "about" allows for small deviations (10% or less) due to non-uniform process conditions and other small deviations from desired process parameters.

[0066] The array of the first and second preferred embodiments is very suitable for vertical stacking of horizontal planar arrays to form a three dimensional array of device levels, each device level containing an array TFT EEPROMs described above. Figure 14 illustrates a three dimensional memory array 200 of the third preferred embodiment containing a plurality of device levels 202, the device levels containing the array 1 or 100 of TFT EEPROMs described above.

[0067] Each device level 202 of the array 200 is separated and decoupled in the vertical direction by an interlayer insulating layer 203. The interlayer insulating layer 203 also isolates adjacent word lines 53 and adjacent portions of the active areas 5 below the respective word lines in each device level 202. Connection between the bit lines, word lines and peripheral or driver circuits in the substrate 206 is made through vertical interlevel interconnects 208.

[0068] Figures 15 through 18 illustrate a method of forming vertical interlevel interconnects 208 between an upper device level 202 and a lower device level 202 or driver circuits in the substrate 206, according to

a fourth preferred embodiment of the present invention. A photoresist mask 209 is formed by applying a photoresist layer over the first N+ polysilicon layer of the control gate layer 33, exposing the photoresist layer through an interlevel interconnect mask and developing the photoresist, as shown in Figure 15. The photoresist mask 209 is used to anisotropically etch the first N+ polysilicon layer 33, the control gate dielectric layer 31, the floating gate rails 9, the tunnel dielectric layer 7, and the active layer 5, stopping the etch on the interlayer insulating layer 203 and the intergate insulating layer 25. Then, the insulating layers 203 and 25 are etched anisotropically using mask 209 to form at least one via extending to the lower device level 202 or substrate 206, as shown in Figure 16. The use of the intergate insulating layer 25 as an etch stop allows the formation of a stepped via, as shown on the left side of Figure 16. The photoresist mask 209 is subsequently removed. Next, a conducting layer 211, comprising a metal layer, such as Ti, W, etc., or a silicide layer, such as TiSi, WSi, etc., is deposited conformally on the first heavily doped N+ polysilicon layer 33, followed by conformal deposition of an optional second N+ polysilicon layer 213. Of course P+ polysilicon may be used instead of N+ polysilicon for these layers. Next, the wordline photoresist mask 35 is formed over layer 213, as shown in Figure 17. Then, as shown in Figure 18, the control gate etch process is performed similar to that shown in Figure 8, to form the plurality of the control gates 43, the plurality of the floating gates 49, and the plurality of the channel island regions 19, as described above with respect to the first embodiment. The vertical interlevel interconnects 208 are formed, as shown in Figure 18. The patterned conducting layer 211 and the second polysilicon layer 213 form the interconnects 208 as well as the upper portion of the control gates 43 (i.e., an upper portion of the word lines or gate lines 53). Thus, at least a portion of the word or gate line comprises the same layer(s) as the interlevel interconnects.

[0069] Preferably, the array of nonvolatile memory devices 200 comprises a monolithic three dimensional array of memory devices. The term "monolithic" means that layers of each level of the array 200 were directly deposited on the layers of each underlying level of the array. A first array or device level 202 of TFT EEPROMs is provided over the substrate 206. The interlayer insulating layer 203 is formed over this array. Then, at least one or more additional arrays or device levels 202 of TFT EEPROMs are monolithically formed on the interlayer insulating layer 203.

[0070] Alternatively, two dimensional arrays may be formed separately and then packaged together to form a non-monolithic three dimensional memory array 200. A plurality of arrays 1 or 100 of TFT EEPROMs are formed on different silicon-on-insulator substrates. The substrates are thinned by polishing or etching the back sides of the substrates. The arrays are then attached or glued to each other to form a three dimensional memory array 200.

[0071] Preferably, the TFTs in a plurality of the levels 202 of the three dimensional array 200 of Figure 14 undergo a recrystallization and/or a dopant activation step at the same time. This reduces the device fabrication time and cost. Furthermore, if each level of the array is subjected to a separate crystallization and/or dopant activation annealing, then the lower levels would undergo more annealing steps than the upper levels. This may lead to device non uniformity because the grain size may be larger in the active areas of the lower levels and/or the source and drain regions may have a different dopant distribution in the lower levels than in the upper levels.

[0072] Each cell in one level 202 of the memory array 200 can be formed using only two photolithographic masking steps. However,

additional masking steps may be needed to form contacts to the bit lines and the word lines. The preferred aspects of the present invention may also be applied to nonvolatile flash memory architectures other than VGA, DuSNOR and SSL-NOR memory. Furthermore, the present invention is not limited to TFT EEPROM flash memory arrays, and also encompasses other semiconductor devices within its scope. For example, the self-aligned transistors may be MOSFETs in a bulk substrate. These self-aligned transistors may be used as non-flash EEPROMs (i.e., EEPROMs where each transistor is erased separately), UV erasable PROMs (EPROMs), mask ROMs, dynamic random access memories (DRAMs), liquid crystal displays (LCDs), field programmable gate arrays (FPGA) and microprocessors.

[0073] The foregoing description of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed, and modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. The drawings and description were chosen in order to explain the principles of the invention and its practical application. The drawings are not necessarily to scale and illustrate the memory array in schematic block format. It is intended that the scope of the invention be defined by the claims appended hereto, and their equivalents.